Prediction of protein-protein interactions using one-class classification methods and integrating diverse biological data

José A. Reyes[1,2,*] and David Gilbert[1]

[1]Bioinformatics Research Centre, Department of Computing Science, University of Glasgow.
[2]Facultad de Ingeniería, Universidad de Talca, Chile.

Summary

This research addresses the problem of prediction of protein-protein interactions (PPI) when integrating diverse kinds of biological information. This task has been commonly viewed as a binary classification problem (whether any two proteins do or do not interact) and several different machine learning techniques have been employed to solve this task. However the nature of the data creates two major problems which can affect results. These are firstly imbalanced class problems due to the number of positive examples (pairs of proteins which really interact) being much smaller than the number of negative ones. Secondly the selection of negative examples can be based on some unreliable assumptions which could introduce some bias in the classification results.
Here we propose the use of one-class classification (OCC) methods to deal with the task of prediction of PPI. OCC methods utilise examples of just one class to generate a predictive model which consequently is independent of the kind of negative examples selected; additionally these approaches are known to cope with imbalanced class problems. We have designed and carried out a performance evaluation study of several OCC methods for this task, and have found that the Parzen density estimation approach outperforms the rest. We also undertook a comparative performance evaluation between the Parzen OCC method and several conventional learning techniques, considering different scenarios, for example varying the number of negative examples used for training purposes. We found that the Parzen OCC method in general performs competitively with traditional approaches and in many situations outperforms them. Finally we evaluated the ability of the Parzen OCC approach to predict new potential PPI targets, and validated these results by searching for biological evidence in the literature.

1 Introduction

The prediction of protein-protein interactions (PPI) has emerged recently as an important problem in the fields of Bioinformatics and Systems Biology, due the fact that most essential cellular processes are mediated by these kind of interactions. High-throughput methods for the direct identification of PPI have been developed including yeast two-hybrid screens (Y2H) [1, 2] and mass spectrometry methods for protein complex identification [3, 4]. Even though high-throughput techniques can increase the number of predicted PPI, in general the data obtained by these methods is often incomplete and suffers from high false-positive and false-negative rates [5]. In order to improve the accuracy and trustability of predicted protein interacting pairs, various studies have previously been developed based on the integrative learning analysis

*Corresponding author: jareyes@dcs.gla.ac.uk

of diverse biological sources of information. These have demonstrated that the combined use of direct and indirect biological insights can improve the quality of predictive PPI models.

The prediction of PPI has been commonly viewed as a classical binary classification problem where the aim is to predict whether any two proteins do or do not interact. Several traditional machine learning methods have been employed in the past for this specific task [6, 7, 8, 9, 10, 11, 12]. These methods generally use supervised learning algorithms where the final objective is to generate a classification model from a gold standard reference set of positive (truly interacting protein pairs) and negative examples (non-interacting pairs). Two main drawbacks have been identified regarding these previous approaches:

i) In general they face a highly imbalanced classification problem, where the number of positive examples is much smaller than the number of negative examples. This affects the quality of the predictive models which may be biased towards the majority class and consequently the minority class examples are poorly predicted. Under-sampling and cost sensitive strategies have been used to deal with the imbalanced problem in some of these previous works whilst others did not report any action about it.

ii) Although the selection of positive examples is based on trustable experimental techniques (i.e. small scale experiments), there is no experimental method to find pairs of proteins which do not interact (negative examples). Therefore certain assumptions have to be made in order to construct a negative gold standard set, which can introduce some bias into the learning process and consequently produces a significant effect on the performance of the classification approach [13].

We propose the use of one-class classification (OCC) methods in this research as a possible solution to these issues. The aim of OCC is to use feature information from only one of the classes, positive examples in this case, to generate a classification model. OCC methods are known to be able to deal efficiently with highly imbalanced classification problems [14]. Additionally, unlike conventional binary classifiers, OCC methods produce classification models which are independent of the kind of negative gold standard set employed.

In this paper we present the results of the application and evaluation of various OCC methods for the prediction of PPI, based on the integrative learning analysis of diverse biological data. Additionally we have carried out a comparative evaluation between OCC methods and several common binary classification techniques, which have been previously employed for this specific task. Previous studies have performed a comparative evaluation between several learning classifiers [7, 12], but did not consider OCC methods as an alternative. We also note that their results are difficult to compare because they have been generated using different reference gold standard sets and biological information. In the next section we present the main methodological aspects considered in this research, followed by a detailed description and analysis of the results obtained. Finally we give some important conclusions.

2 Methods

2.1 One-class classification methods

The common issue of OCC problems is that feature information is available for only one of the classes, called the *target class*, and this is employed to generate a classification model. The

OCC model is constructed with the aim of characterizing and describing the target examples, and afterwards it is used to distinguish target examples from all other examples which have been classified into a single different category called the *outlier class*. The general task in OCC can be regarded as being similar to conventional binary classification methods, in that a decision boundary or separation model is used to separate examples of the two classes (target and outliers). However OCC methods face a harder task because the decision boundary is mainly supported by examples of the target class and hence less information is employed to build and validate it. Consequently a sufficiently representative sample of target examples is needed to generate a more accurate descriptive model in order to improve the OCC performance.

In this research we consider the task of prediction of PPI as a OCC problem, in the sense that only examples of one class (positive interaction examples) are available and/or trustable, becoming the *target class*. The resulting classification model is independent of the kind and quality of the set of negative examples employed because the OCC approach is mainly based on the description of examples of the target class; this could potentially solve the problem of trustability associated with the selection of the negative class. In order to develop a comparative performance evaluation between OCC and conventional classification methods, a set of negatives examples should be selected as the *outlier class*. This is because it is necessary to use examples of both classes for training and testing purposes when considering considering conventional binary classifiers. Under these conditions the performance of OCC methods can be evaluated in a manner similar to that for conventional binary classification techniques, by estimating the misclassification error, i.e the target class error (or false-negative rate), and when outlier examples are available, the outlier class error (or false-positive rate).

OCC methods can be classified according to the way in which they analyze, describe and generate a model for the separation of targets and outlier examples [15]. Here we consider two types, as follows. *(A) Density estimation* methods based on the estimation of the probability density distribution of the training data using some probabilistic model (i.e. Gaussian distribution). A threshold is selected and then used to compare with the density of new objects in order to classify them. *(B) Boundary* methods based on the generation of a frontier or boundary around the target objects, which is optimized to accept most of the target examples and at the same time reject most of the outliers. Four different OCC learning approaches were evaluated in this research, namely three density estimation methods (single Gaussian estimation, mixture of Gaussian and Parzen density estimation) and a boundary approach (Support vector data description SVDD). The dd_tools Matlab toolbox (`http://www-ict.ewi.tudelft.nl/~davidt/dd_tools.html`) was utilized to develop the experiments associated with the application and evaluation of all OCC methods. A detailed explanation of each of the OCC methods employed is presented in the supplementary material.

2.2 Reference data set

In this research we focused on the prediction of co-complexed proteins pairs (pairs of proteins which are co-members of the same complex). In order to evaluate different machine learning methods we need a reference data set (gold standard) containing positive and negatives examples. We used the same gold standard sets employed by Lin et al. [7] for the study of PPI in yeast. These comprise 2,104 positive examples (true interacting protein pairs) derived from the MIPS complex catalogue [16] and 172,409 negative examples (non-interacting protein pairs) related to protein pairs where the members are localized in different cell compartments and

consequently are likely not to interact between them. This reference data set is a subset of that used by Jansen et al. [6], considering only examples where complete information for each one of the biological features is available.

2.3 Biological features

An important motivation for this research is that the integration of diverse kinds of biological data/information could potentially improve our ability to predict protein-protein interactions. Four different types of biological information were considered following [6] and [7]:

m-RNA expression, following the assumption that proteins which are members of the same complex are commonly expressed simultaneously. The Pearson correlation was estimated for every protein pair considering two different well known studies: the Rosetta compendium [17] and cell cycle time series analysis [18], generating two numeric values between -1 and 1 which are incorporated as features.

functional similarity of protein pairs was estimated from the gene ontology (GO) [19] and the MIPS [16] functional catalog, obtaining two new numeric features. The assumption here is that proteins in the same complex tend to participate in the same biological processes.

Essentiality information [16], assuming that two proteins in the same complex are essential or non essential for cell survival. This feature is then characterized by three possible categories (i.e. both proteins are essential or both are non-essential or only one of them is essential), and is represented by a three dimensional vector taking discrete values of +1 or -1 according to each case.

High-throughput experimental interaction data from Y2H and mass spectrometry based experiments were integrated as features. Four different experimental studies have been considered [1, 2, 3, 4]. In each case a discrete value of +1 or -1 is assigned to indicate whether the components of a protein pair do interact or do not interact respectively.

Numerical features were normalized to obtain a distribution with a mean of 0 and standard deviation of 1, in order to to put all data in the same range of values and to avoid possible numerical difficulties associated with imbalanced ranges. Every pair of proteins available in the reference data set was represented by a 11-dimensional vector X_i containing the information for the biological features considered here, and a label Y_i which can take two values depending on whether each of the proteins pairs do really interact ($Y_i = 1$) or not ($Y_i = -1$).

2.4 Conventional machine learning methods

A representative group of conventional or traditional machine learning techniques, which have been previously used for the task of prediction of PPI, was selected in order to undertake a comparative performance evaluation with OCC methods for this specific task. These includes: Decision Trees (DT), Naive Bayes (NB), Logistic Regression (LR) and Support Vector Machines (SVM). The WEKA machine learning library [20] was used to perform the experiments related to DT, NB and LR, while the evaluation of SVM was carried out using the MATLAB interface to the SVM-light toolbox (`http://svmlight.joachims.org`).

2.5 Performance evaluation

OCC and conventional learning approaches were evaluated in different training/testing scenarios varying, for instance, the number of negative examples used to train each of the models. A ten-fold cross validation procedure was carried out for every evaluation in order to assess the variability of the models generated. In each situation the negative examples which were not utilized in the training step were also included in the testing evaluation. This testing strategy differs from previous approaches used for this task, where only a fraction or sub-sample of the negative gold standard examples was considered to test the models. We think that by including all the available putative negative information each time we test our models, we are carrying out a more relevant and at the same time more challenging evaluation for the prediction of PPI.

Receiver Operator Characteristic (ROC) curves, illustrating the tradeoff between the false-positive rates and true-positive rates, were generated for each approach under the different scenarios evaluated. The area under the ROC curve (AUC) was calculated for each case to evaluate the overall performance of different learning algorithms. AUC scores seem to be a better evaluation measure than simple accuracy in imbalanced class problems [21].

We also calculated partial AUC scores, which are related to the normalised area under a fraction of the whole ROC curve which represents a condition of special interest. For example in the situation of severe class imbalance it seems more relevant to evaluate the performance in the region of low values of false-positive rates [22], which is the case in the prediction of PPI tasks. In our approach we are interested in evaluating and comparing the performance of the different classifiers under conditions of a low false-positive rate. The aim of this is to maximise the number of real interacting protein pairs predicted while minimizing the number of false-positive predicted ones. This is of especial interest for biologists working in the identification and validation of new PPI, because they can focus on the study of only the top ranked predicted PPI targets instead of evaluating many randomly selected protein pairs. We considered the area under the ROC curve up to the first 50 false-positive examples (AUC-50), which has become a commonly accepted performance measure for this specific task [11, 12].

Mean values and standard deviation for AUC and AUC-50 were calculated based on the ten fold cross-validation individual results, in order to compare the performance of different approaches. When the difference was unclear between the AUC or AUC-50 values for two methods, the Wilcoxon signed rank statistical test [23] for the median of the differences between them was computed considering a 5% significance level, in order to obtain stronger evidence that one of the methods performed better than the other.

3 Results

3.1 Evaluation of diverse OCC methods

Four different OCC methods were used for the problem of PPI prediction including: Gaussian density estimation, Mixture of Gaussian density estimation, Parzen density estimation and Support Vector Data Description (SVDD). The methods were evaluated on a balanced class set using all the positive examples available and an equal size sample of negative examples randomly selected from the whole negative gold standard set. This was done because some of

the OCC methods can take advantage of the use of a sample of negative examples to improve their performance. This procedure was repeated ten times using diverse sub-samples of negative pairs. The results of the estimation of AUC and AUC-50 scores for the OCC performance evaluation are shown in Table 1 where the mean and standard deviation are given.

Table 1: Comparison of AUC and AUC-50 values for different learning methods evaluated

Method	AUC	AUC-50
OCC methods:		
SVDD	0.9768 ± 0.0033	0.2455 ± 0.0325
Gaussian	0.9377 ± 0.0136	0.1224 ± 0.0136
Mixture of Gaussian	0.9853 ± 0.0096	0.2269 ± 0.0513
Parzen	0.9801 ± 0.0075	0.4010 ± 0.0282
Conventional methods:		
Decision trees (DT)	0.9946 ± 0.0033	0.2129 ± 0.1903
Naive Bayes (NB)	0.9908 ± 0.0017	0.2299 ± 0.0275
Logistic Regression (LR)	0.9928 ± 0.0018	0.0917 ± 0.0307
Support Vector Machines (SVM)	0.9939 ± 0.0016	0.2687 ± 0.0250

The results for the global AUC scores show that there is no significant difference between most of the OCC methods evaluated, with the exception of the simple Gaussian density estimation method which exhibits the lowest overall performance. On the contrary, the analysis of the results for the AUC-50 scores clearly shows that the Parzen density estimation method (AUC-50 = 0.401) by far outperforms the rest of the OCC methods considered here. The good performance obtained by the Parzen method can be explained because this density estimation method takes into account the information of every target example available. This is different to the rest of the OCC approaches evaluated, where for example only an average probability density estimation from the available data is employed as in the case of Gaussian and Mixture of Gaussian approaches, or in the case of SVDD method where just few examples are utilised to support a boundary between target and outlier examples.

The second best performance for OCC methods considering AUC-50 scores is obtained by the SVDD approach using a Gaussian kernel (AUC-50 = 0.2455). We note that a recent paper by Alashwal et al. [24] used one-class support vector machines (OCSVM) [25], which is an extension of the classical binary SVM technique, to deal with the task of prediction of PPI. In that work the authors only considered one biological feature based on protein sequence and domain information, reporting that the best results are obtained using a Gaussian kernel. In contrast, in our research we evaluated several different OCC approaches, used diverse biological features and also carried out a comparative performance evaluation with several conventional binary classification methods. Moreover, it has been shown that the SVDD and OCSVM techniques give equivalent solutions [15, 25] when using a Gaussian kernel.

3.2 Comparative evaluation between OCC and conventional classifiers

The Parzen OCC method was selected, due to its good performance, to be compared in a more exhaustive evaluation with several conventional classifiers such as Decision Trees (DT), Naive Bayes (NB), Logistic Regression (LR) and Support Vector Machines (SVM). Firstly all the

learning approaches were evaluated on the same ten different balanced class sets previously used. Estimates for AUC and AUC-50 scores for these experiments are given in Table 1.

Comparative analysis of overall AUC scores shows that conventional classifiers perform only slightly better that the Parzen OCC approach. This was expected because the task associated with OCC only uses examples of one class to generate a classification model. However in relation to the AUC-50 comparative evaluation, we found that the Parzen OCC approach clearly outperforms all conventional classification techniques (AUC-50 = 0.401). The performance of conventional classifiers in these cases is only comparable with some of the other OCC methods previously evaluated, and sometimes worse as the case of the LR approach. SVM showed the best performance for the conventional classifiers (AUC-50 = 0.2687). It is interesting to note that DT exhibits high variability compared with the rest of the methods evaluated. The detailed analysis of AUC-50 results shows that in some of the ten fold cross validation subsets DT performs better than OCC methods but in others (the majority) it performs very poorly. The Wilcoxon signed rank test [23] was applied in this case showing that effectively the Parzen OCC method outperforms the rest of conventional classifiers.

The difference between the AUC and AUC-50 analysis can be clearly appreciated from the ROC curves of the different learning methods evaluated. Figure 1(a) shows an example of the ROC curves for the different learning techniques used in the evaluation of one cross validation subset. No important differences between these ROC curves is observed and consequently there is no significant difference in total AUC scores. When we focus on the portion of these curves related to the AUC-50 region, presented in Figure 1(b), there are clear differences in the performance of the diverse methods. In this region the Parzen OCC method outperformed the rest of the conventional learning approaches evaluated. This is still the case if we extend the partial AUC analysis up to the first 100 false-positive examples. This corroborates our assumption that analysis based on partial AUC scores (i.e. AUC-50) is more appropriate than that using overall AUC scores, for predicting PPI.

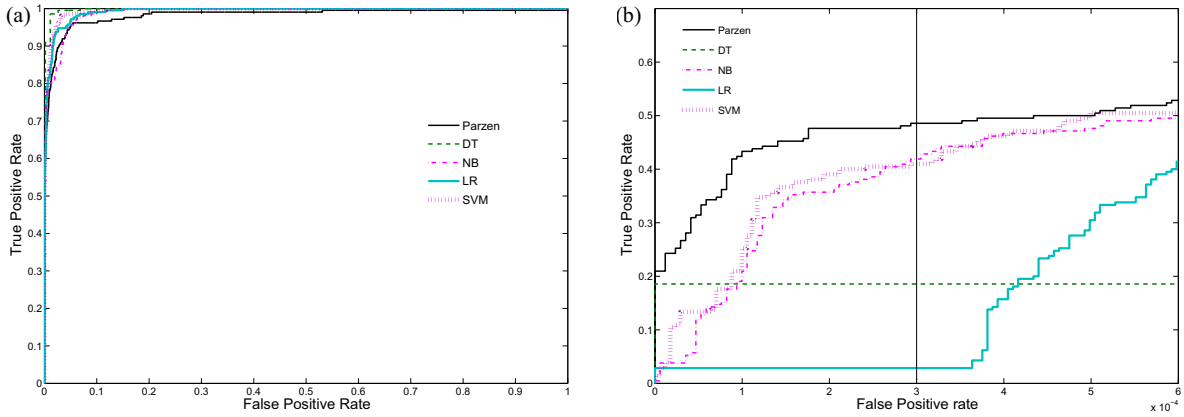

Figure 1: Example of ROC curve analysis: (a) Whole ROC curves for the different learning methods evaluated. (b) Partial ROC curves for the different learning methods evaluated. The vertical line indicates the point where approximately the first 50 false-positive examples are reached.

3.3 Comparative evaluation on different scenarios

We also evaluated and compared the effect of the use of negative examples in the performance of the diverse learning approaches. Different scenarios were generated varying the number

of negative examples used for training the respective models, from none to all the negative examples available. Figure 2 shows the performance results, measured as AUC-50 scores, for all the situations considered.

Firstly we analysed the cases where less negative than positive examples were used to train the models, including the balanced class scenario when 2,104 negative examples are employed. The Parzen OCC method clearly outperforms the rest of conventional learning techniques, exhibiting a very stable performance in the different situations. This can be explained because it only uses positive examples for training purposes. On the contrary, the performance of conventional classifiers tends to decrease as less negative information is used. SVM exhibits the best performance for binary classifiers followed by the NB approach. DT and LR exhibit low performance and high variability compared with the rest of methods evaluated. Note that in the situation where no negative examples are used, only the Parzen OCC method can be employed and consequently no results for conventional classifiers are available.

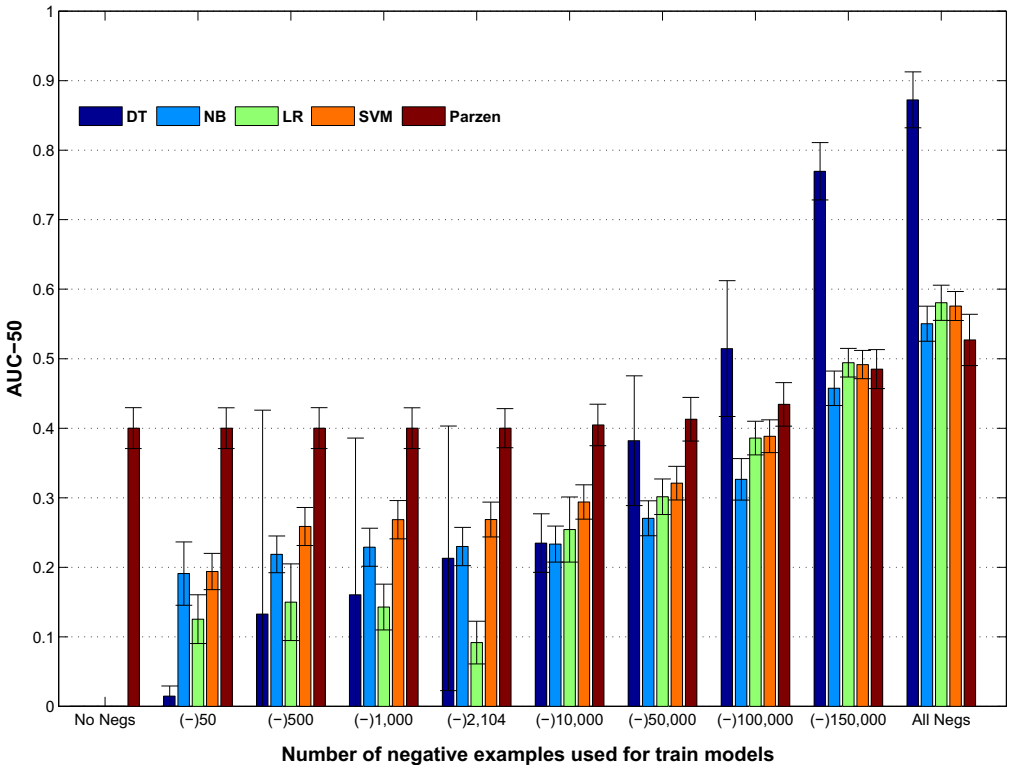

Figure 2: AUC-50 comparison for different learning methods evaluated, showing the effect of reducing and incrementing the number of negative examples used to train the models. The balanced class scenario is when 2,104 negative examples are used for training. Note that no corrective action was taken for any of the imbalanced class situations.

The analysis is quite different for scenarios where more negative than positive examples are employed to train the models. The Parzen density estimation OCC technique tends to maintain its performance stability and a significant increment in the AUC-50 performance only occurs when more than 50,000 negative examples are employed. This can be explained because in these cases the models were tested on a reduced number of negative examples (most of the negative information is used to train the models). The performance of conventional classifiers tends to increase gradually as more negative examples are incorporated for the generation of their respective classification models. This was expected because these techniques can take advantage of the negative object class information.

The Parzen OCC method performs very competitively in most of the scenarios evaluated, and outperforms the other methods up to the case where 50,000 negative examples are used for training. At this point the DT technique performs as well as the Parzen OCC approach. Thereafter the DT method outperforms all the rest of the learning approaches, suggesting that DT is the traditional binary learning approach most influenced by the availability of the negative class information. Other conventional classifiers evaluated (NB, LR and SVM) do not exhibit outstanding performance and slightly outperform the Parzen OCC method only when all available negative examples are used.

Finally we studied the effect of imbalanced classes on the performance of the different classifiers. While OCC methods are intrinsically able to cope with this situation, this is not the case for conventional classifiers. Consequently some strategy is needed to deal with the imbalanced class problem. Here we used a cost-sensitive analysis, where the misclassification cost for examples of the minority class is bigger than the misclassification cost for the majority class (note that on the different scenarios the minority class is not always the same). In situations where fewer negative than positive examples were used, we observe an increment in the performance of most of the conventional classifiers, reaching AUC-50 scores similar to those obtained for each approach in the balanced class scenario. The exception is the NB approach, the performance of which was almost invariant in these cases. When more negative than positive examples were used, the AUC-50 performance for all conventional classifiers tended to decrease in comparison with those obtained without cost-sensitive analysis. This can be explained because in these cases the classification model is generated considering positive and negative examples' information in a balanced way and is not biased towards negative class information. Another accepted strategy to deal with the imbalanced class problem is to under-sample the majority class; we have done this when training on ten different balanced class sets (see section 3.2).

The analysis of the results presented in this section strongly suggests that the performance of conventional binary classification models is highly affected by the presence or absence of negative examples. This can also explain the high performance (AUC-50) observed for conventional classifiers when all negative examples are employed for training. Another explanation for this observed high performance is the availability of a high-quality negative gold standard set (protein pairs located on different cell localization), which has been previously discussed in [11] and [13]. However this will not be the case when undertaking the prediction of PPI on other organisms when protein cell localization information is unavailable.

3.4 Evaluation of biological feature importance

We evaluated the individual effect of the different biological features used in this research on the performance of the Parzen OCC approach. For this we removed each of the biological attributes one at time from the data set and tested the effect of this action on the AUC and AUC-50 scores, compared with those obtained when all available biological information is used. Table 2 shows the results of this procedure.

The major effect on the Parzen OCC performance occurs when either functional similarity or m-RNA expression data are removed. This is consistent with results previously reported in the literature [7, 9, 12]. It is interesting to observe that the overall AUC performance only increases when high-throughput information is removed, which can be explained due the high false-positive and false-negative rates associated with these kinds of features.

Table 2: Evaluation of the individual effect of the different biological attributes in the performance of the OCC parzen approach

Feature description	AUC	AUC-50
ALL features	0.9801 ± 0.0075	0.4010 ± 0.0282
GO removed	0.9186 ± 0.0121	0.2094 ± 0.0189
MIPS removed	0.9412 ± 0.0135	0.1983 ± 0.0225
m-RNA expression removed	0.9775 ± 0.0050	0.1883 ± 0.0238
Essentiality removed	0.9800 ± 0.0081	0.3380 ± 0.0273
High-throughput removed	0.9887 ± 0.0037	0.3463 ± 0.0261

3.5 Prediction of new potential PPI targets using Parzen OCC method

Finally we evaluated the ability of the Parzen OCC approach to predict new potential PPI, which could be used as a targets in future investigations. For this we generated a new set of random protein pairs which were not included on our positive and negative gold standards sets. We were able to collect a set of approximately 518,000 protein pairs examples with complete biological information from the data previously used in [6]. We classified the examples in the random set using the Parzen OCC model trained on all positive examples available (parameters being optimized on ten fold cross validation procedure), and found that 928 of them were predicted as a new potential PPI.

We focused on the analysis of the top 50 new potential PPI with the highest prediction scores generated by the Parzen OCC model. This score is the probability associated with the the positive examples class and consequently can be seen as a confidence value. To validate our predictions we employed the INTACT database (`http://www.ebi.ac.uk/intact`), which compiles molecular interactions reported in published literature, containing information for around 50,000 binary protein interactions for yeast (May 2006). We found that of the 50 top ranked examples, 36 were supported by at least one reference in INTACT. These were mostly associated with mass spectrometry experiments which are related with the identification of groups of proteins that interact to form complexes. This is statistically significant considering that if we randomly selected 50 protein pairs not in the positive gold standard, the probability that 36 of them will be annotated in INTACT is very low ($p < 10^{-77}$) using Fisher's exact test [26]. The list of the top 50 potential new PPI targets is given in the supplementary material.

4 Conclusions

The research described in this paper has focused on the application and evaluation of one-class classification (OCC) methods for the problem of prediction of protein-protein interaction (PPI). We also considered the use of diverse biological data types in order to develop a joint integrative learning analysis.

Among various OCC methods evaluated, the Parzen OCC density estimation approach clearly exhibited the best performance. This can be explained because the Parzen OCC technique utilises all examples in the training set to generate a classification model unlike the other OCC methods investigated here. This approach was then selected to develop a comparative perfor-

mance evaluation against several well known conventional machine learning methods. Different scenarios were considered varying the number of negative examples used to train the models. We found that the Parzen OCC approach performs very competitively and outperforms the rest of conventional classifiers in most of the situations up to the case where the ratio of negative to positive examples is approximately 25 to 1.

We have demonstrated that for this specific task, the performance of conventional binary classification approaches is highly influenced by the quantity of negative examples used to train the respective models. This suggests that classification models generated from these type of methods are more reliant on negative information (in this case an untrustworthy set of negative PPI examples) than on positive information (experimentally corroborated PPI examples).

Our results indicate that the task of the prediction of PPI can indeed be formulated as an OCC problem where the predictive model is based on real (trustworthy) PPI data. In the specific case of prediction of co-complexed proteins we found that the Parzen OCC method is able to generate models which perform competitively with those generated by conventional classifiers, independently of the quality and quantity of the negative examples available. We have also carried out an initial study about the ability of the Parzen OCC approach to predict new potential PPI targets, showing that many of the highly ranked new predictions can be validated by reference to published results in the literature.

Acknowledgements

This work has been supported by the Programme Alβan, the European Union Programme of High level Scholarships for Latin America, scholarship E04D034854CL.

References

[1] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. Nature, 403(6770):623–627, 2000.

[2] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci, 98(8):4569–4574, 2001.

[3] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 415(6868):141–7, 2002.

[4] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, et al. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. Nature, 415(6868):180–183, 2002.

[5] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, et al. Comparative assessment of large-scale data sets of protein-protein interactions. Nature, 417(6887):399–403, May 2002.

[6] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, et al. A bayesian networks approach for predicting protein-protein interactions from genomic data. Science, 302(5644):449–453, 2003.

[7] N. Lin, B. Wu, R. Jansen, M. Gerstein, and H. Zhao. Information assessment on predicting protein-protein interactions. BMC Bioinformatics, 5(1):154, 2004.

[8] L. Zhang, S. Wong, O. King, and F. Roth. Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinformatics, 5(1):38, 2004.

[9] L. J. Lu, Y. Xia, A. Paccanaro, H. Yu, and M. Gerstein. Assessing the limits of genomic data integration for predicting protein networks. Genome Res., 15(7):945–953, 2005.

[10] Y. Qi, J. Klein-Seetharaman, and Z. Bar-Joseph. Random forest similarity for protein-protein interaction prediction from multiple sources. In Pacific Symposium on Biocomputing, 2005.

[11] A. Ben-Hur and W. S. Noble. Kernel methods for predicting protein-protein interactions. Bioinformatics, 21(suppl 1):i38–i46, 2005.

[12] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins: Structure, Function, and Bioinformatics, 63(3):490–500, 2006.

[13] A. Ben-Hur and W. S. Noble. Choosing negative examples for the prediction of protein-protein interactions. BMC Bioinformatics, 7(Suppl 1):S2, 2006.

[14] N. V. Chawla, N. Japkowicz, and A. Kotcz. Editorial: special issue on learning from imbalanced data sets. SIGKDD Explorations, 6(1):1–6, 2004.

[15] D. M. J. Tax and R. P. W. Duin. Support vector data description. Machine Learning, 54(1):45–66, 2004.

[16] H. W. Mewes, D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, et al. Mips: a database for genomes and protein sequences. Nucl. Acids Res., 30(1):31–34, 2002.

[17] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, et al. Functional discovery via a compendium of expression profiles. Cell, 102(1):109–126, July 2000.

[18] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, et al. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell, 2(1):65–73, July 1998.

[19] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 25(1):25–29, May 2000.

[20] I. H. Witten and E. Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

[21] J. Huang and C. X. Ling. Using auc and accuracy in evaluating learning algorithms. IEEE Trans. Knowl. Data Eng., 17(3):299–310, 2005.

[22] C. Drummond and R. C. Holte. Learning to live with false alarms. In Workshop on Data Mining Methods for Anomaly Detection. Eleventh ACM SIGKDD, August 2005.

[23] F. Wilcoxon. Individual comparisons by ranking methods. Biometrics Bulletin, 1(6):80–83, 1945.

[24] H. Alashwal, S. Deris, and R. M. Othman. One-class support vector machines for protein-protein interactions prediction. International Journal of Biomedical Sciences, 1(2):120–127, 2006.

[25] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. Neural Computation, 13(7):1443–1471, 2001.

[26] R. A. Fisher. On the interpretation of χ^2 from contingency tables and the calculation of p. Journal of the Royal Statistical Society, 85(1):87–94, 1922.

Supplementary material: Prediction of protein-protein interactions using one-class classification methods and integrating diverse biological data

José A. Reyes[1,2] and David Gilbert[1]

[1]Bioinformatics Research Centre, Department of Computing Science, University of Glasgow.
[2]Facultad de Ingeniería, Universidad de Talca, Chile. Contact: jareyes@dcs.gla.ac.uk

S1 - Description of one-class classification (OCC) methods

In this supplementary section we give a detailed description of the four OCC methods evaluated in this research (3 density and 1 boundary approaches). This has been carried out based on the implementation of each of these methods in the dd_tools Matlab toolbox available at (`http://www-ict.ewi.tudelft.nl/~davidt/dd_tools.html`), which was utilized to develop the experiments associated with the application and evaluation of all OCC methods.

Gaussian density estimation:

This is the simplest of the OCC density approaches. The examples of the target class used for training are modeled as a Gaussian distribution. In the dd_tools implementation the complete density estimation is not obtained and just the Mahalanobis distance is employed and calculated for each example X as:

$$f(X) = (X - \mu)^T \sum^{-1} (X - \mu) \tag{1}$$

where the mean μ and the covariance matrix $\sum$ are estimated from the entire sample of objects used. The $f(X)$ value for new objects is then compared against a threshold θ and classified as a target if $f(X) \leq \theta$ or else as an outlier.

Mixture of Gaussian density estimation:

In this case a linear combination of several (i.e. N) different Gaussian distributions is employed to model the target class examples used for training, obtaining a more flexible model compared with the single Gaussian distribution approach. The training data is divided into N different clusters, each of which is modeled by a single Gaussian distribution. The distance function $f(X)$ changes in this case to the form:

$$f(X) = \sum_{i=1}^{N} \alpha_i \exp(-(X - \mu_i)^T \sum_{i}^{-1} (X - \mu_i)) \tag{2}$$

where α_i are the mixing coefficients. The parameters of each cluster μ_i, $\sum_i$, and α_i are optimized using the EM algorithm. A threshold θ is fixed again and used to classify new objects as in the previous case. For this approach it is possible to include outlier objects in the training

phase, setting independent mixtures of Gaussian distributions for both target and outlier examples, considering N_{target} and $N_{outlier}$ different clusters. The number of clusters considered for target and outlier data should be fixed and consequently can be varied in order to obtain an optimal performance of the model.

Parzen density estimation:

In Parzen density estimation an independent Gaussian distribution is considered for each one of the T target objects used to train the model. Consequently in this case the distances to all training objects have to be considered. In the dd_tools implementation of this approach the function $f(X)$ is as follows:

$$f(X) = \sum_{i=1}^{T} \exp(-(X - X_i)^T h^{-2} (X - X_i)) \tag{3}$$

The smoothing parameter h, commonly called the *Parzen width*, is introduced here and is related to the width of a region R (in a Gaussian space) generated around each object in order to separate the target from outlier zones. The rest of the classification process is similar to the previous density approaches. The value of h can be varied in order to optimize the performance of the model.

Support vector data description (SVDD):

This technique is a boundary approach based on the binary Support Vector Machines (SVM) theory. The aim of SVDD is to create a closed hyper-spherically shaped boundary around the target class examples used to train the model. Following the description in [Ref-1, Ref-2] the hyper-sphere is characterized by the centre a and radius R, and is supported for several objects as in the case of SVM. The objective then is to minimize the volume of the sphere which is possible by minimizing the value of R^2. This minimization problem is similar to that in the SVM approach and consequently it is possible to generate the same kind of approximation solution. The SVDD method can also employ a more flexible representation of the data using different kernel functions (i.e. linear, polynomial and Gaussian kernels). This approach permits the use of outlier examples in the training stage in order to generate a tighter description of the hyper-spherical boundary. The kernel type and its respective parameters can be varied in this implementation in order to obtain the optimal performance conditions.

S5 - List of potential new PPI targets

In this supplementary section we list the top 50 new potential PPI targets predicted with the Parzen density OCC approach (Table 1). This table includes the following information: Column 1 enumerates the 50 examples; columns 2 and 3 give the systematic ORF names (i.e. ID-1 and ID-2) for both proteins in each of the new PPI pairs predicted; finally column 4 shows the predictive score (P) for each new predicted PPI which is related to the degree of confidence associated to each pair according to the classification model.

Table 1: List of 50 highly ranked new potential PPI targets predicted by the Parzen OCC method

No	ID-1	ID-2	P
1	YDR025W	YLR029C	0.93420
2	YOL039W	YOL139C	0.93085
3	YBR189W	YOR063W	0.92811
4	YKL156W	YPL131W	0.92766
5	YBR118W	YPL131W	0.92748
6	YML063W	YPL131W	0.92665
7	YKL156W	YPL143W	0.92578
8	YGL135W	YNL178W	0.92496
9	YBR048W	YLR029C	0.92441
10	YHR010W	YNL178W	0.92375
11	YBR189W	YPL143W	0.92253
12	YBL092W	YML063W	0.92183
13	YBR189W	YPL237W	0.92140
14	YEL034W	YOL127W	0.91935
15	YBR189W	YMR260C	0.91858
16	YEL034W	YLR340W	0.91823
17	YDL082W	YNL178W	0.91801
18	YDR382W	YNL178W	0.91736
19	YBR118W	YLR029C	0.91652
20	YKL156W	YNL244C	0.91650
21	YML063W	YNL244C	0.91649
22	YDL136W	YNL244C	0.91628
23	YDL082W	YLR249W	0.91615
24	YGL135W	YHL015W	0.91591
25	YEL034W	YOR063W	0.91568
26	YDL191W	YNL244C	0.91460
27	YDR064W	YGL135W	0.91417
28	YEL034W	YKL060C	0.91353
29	YHL015W	YPL220W	0.91250
30	YBR118W	YOL040C	0.91235
31	YHR010W	YNL244C	0.91169
32	YDR025W	YLR249W	0.91152
33	YBR118W	YOL127W	0.91072
34	YML024W	YNL244C	0.91037
35	YNL244C	YPL220W	0.90981
36	YML024W	YPL143W	0.90814
37	YHR010W	YPL237W	0.90769
38	YER131W	YPL143W	0.90761
39	YER131W	YLR075W	0.90642
40	YBL027W	YHL015W	0.90638
41	YKL156W	YOR063W	0.90594
42	YGL135W	YOL139C	0.90561
43	YBR189W	YLR075W	0.90505
44	YDL130W	YDR064W	0.90369
45	YDL136W	YPL237W	0.90351
46	YDL136W	YNL178W	0.90273
47	YAL003W	YDR418W	0.90200
48	YEL034W	YOL040C	0.90180
49	YDL082W	YOL040C	0.90125
50	YDR385W	YOL040C	0.90094

References

[Ref-1] D. M. J. Tax. One-class classification. PhD thesis, Delft University of Technology, 2001.

[Ref-2] D. M. J. Tax and R. P. W. Duin. Support vector data description. Machine Learning, 54(1):45–66, 2004.